TERABITS

Prototyping Future Networks

Optical Fiber Conference
7-9 March 2006

Naval Research Laboratory

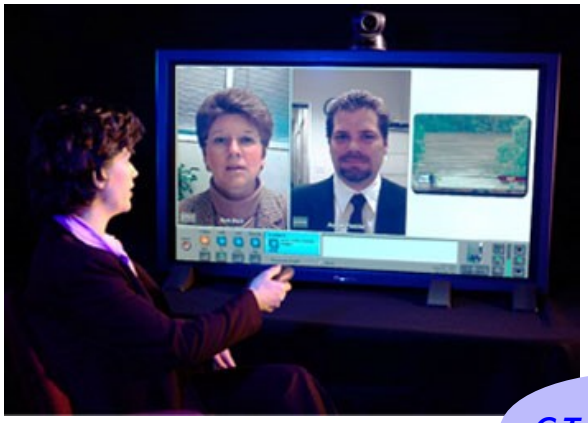
OFC/NFOEC Team …

A Terabit Challenge .

*Build a Global "Large Data" Network Infrastructure to **Rapidly Access** and **Produce Knowledge** from the **Best Information** available from **Federated, Distributed** information assets*

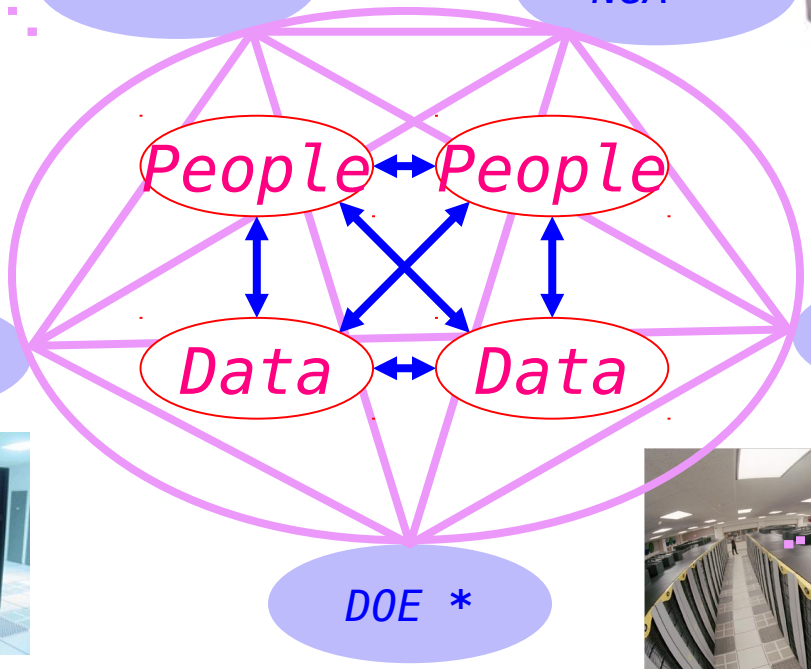
- *Integrate **federated, distributed** computational grids, realtime sensors, and digital historical information*
- *Scale to support **exponentially** increasing data archives*
- *Privacy, authenticity and security demands: **InfoAssured***
- *Affordable … highly available … **E2E QoS/QoP** flows*
- *Legacy and rapidly evolving technology integration*
- *Perf, NetOps, Information Assurance tools/sensors*
- *Reachback, Traceback realtime capabilities*

"Expose interfaces early and ofte



An Enterprise View …

* Hypothetical sites



"DATA CONFERENCING"

… multiple sites, people, P2P seamlessly interacting!

big *fast* "terabytes/hour" data problem …

… efficiently interface high performance optical networks directly to

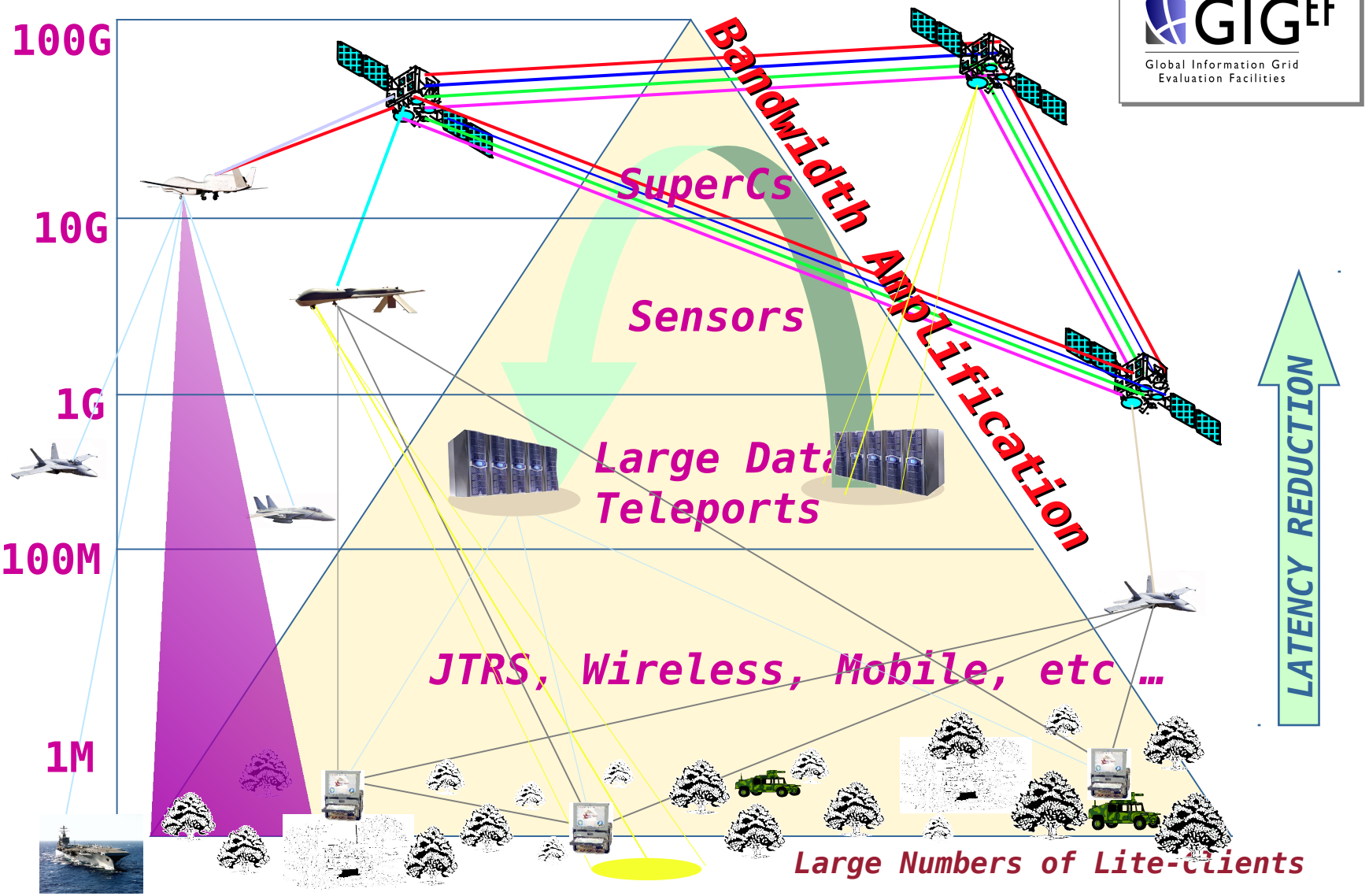
- *Supercomputers*
- *Grid Clusters*
- *Visualization, SuperHDTV*
- *HR Motion Imagery*
- *TSAT Tactical Comms*
- *GIS Imagery/Weather/Oceans*
- *2D/3D workstations*
- *Online Digital Asset Archives*
- *Hyperspectral …40K x 40K*
- *Optical LAN* networks
- *Virtualized Ground Station*

- Interfaces need to scale as networks scale
- Interface programming model and semantics familiar and friendly
- Minimum of equipment required for each *lambda* connection
- WAN transport protocol semantics simply abstracted from applications
- Sustained performance across the WAN approaches *full wire-speed*

-Routinely exchanging multi-TByte streamed data sets long haul during daily workflows from sensors …

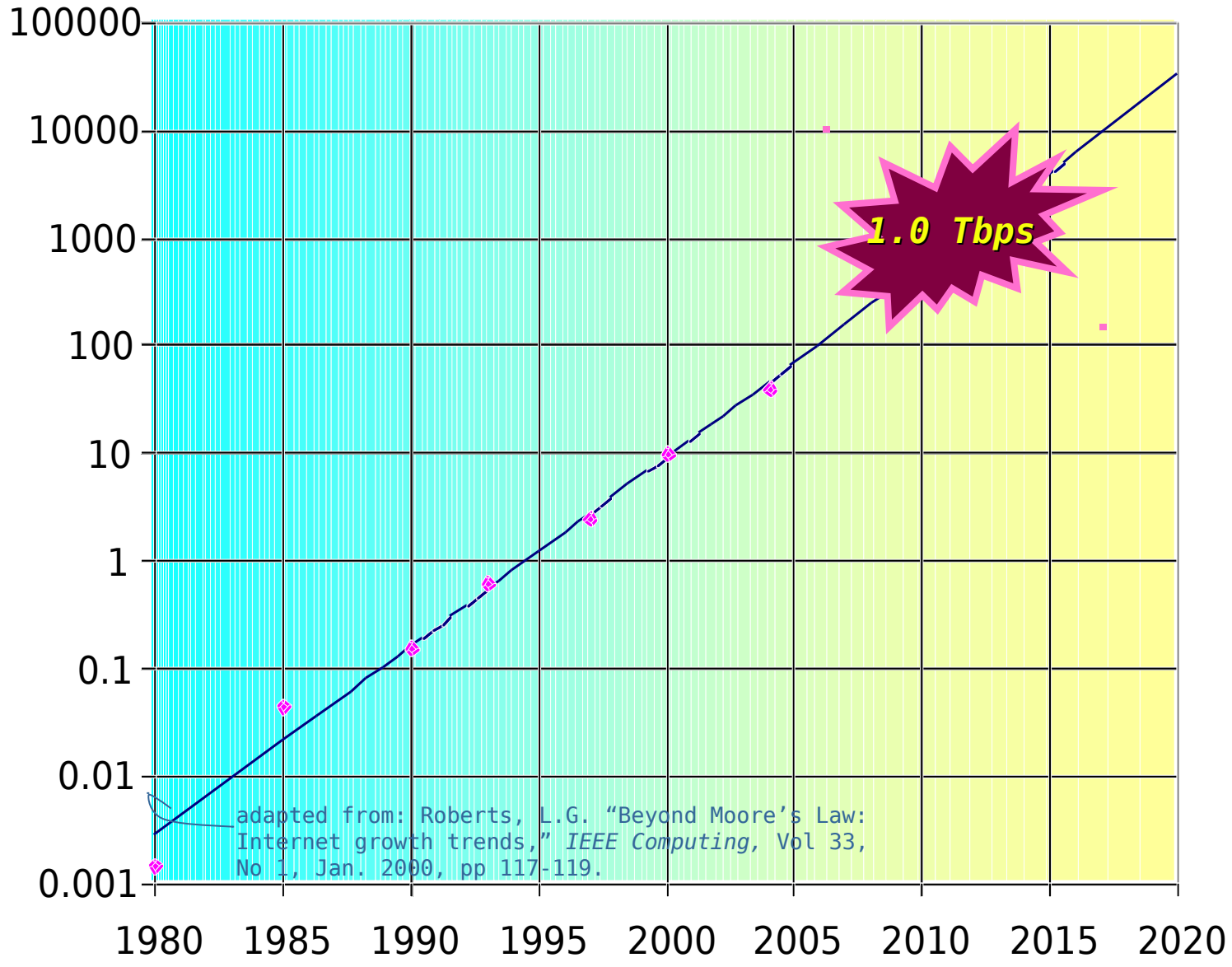
-Multi-PetaByte online distributed, federated archives

A Net-centric Architecture . . .



"… a single packet triggers High Bandwidth Flows …"

Network Growth Trend . . .



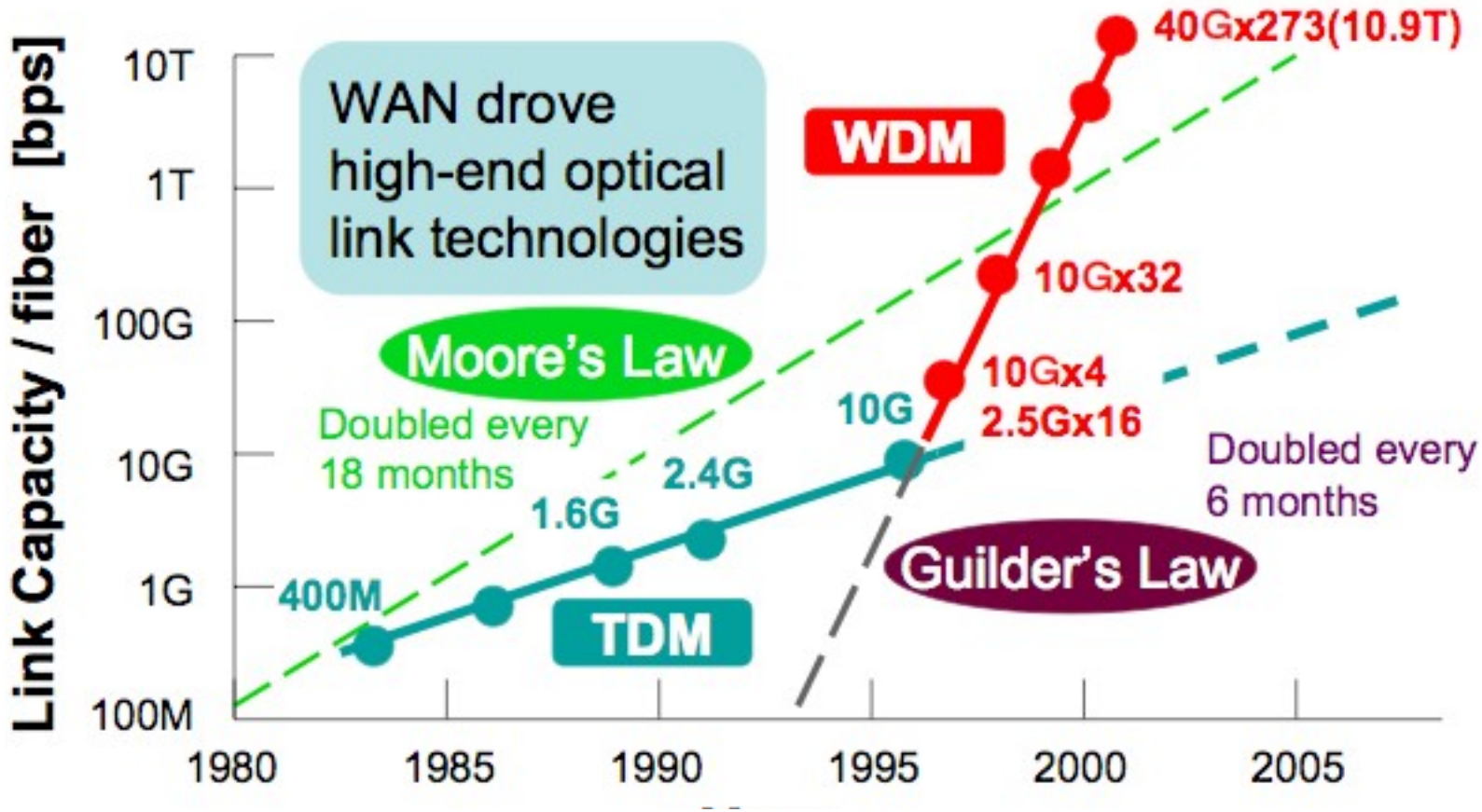
A Few, Familiar Definitions …

Moores' Law: "The capacity of a microprocessor doubles every 18 months"…

Guilder's Law: "Bandwidth will triple every 12 months"… *bandwidth will rise at a rate three times the rate at which processing power is increasing; at the moment processing power is doubling every year and bandwidth every four months.*

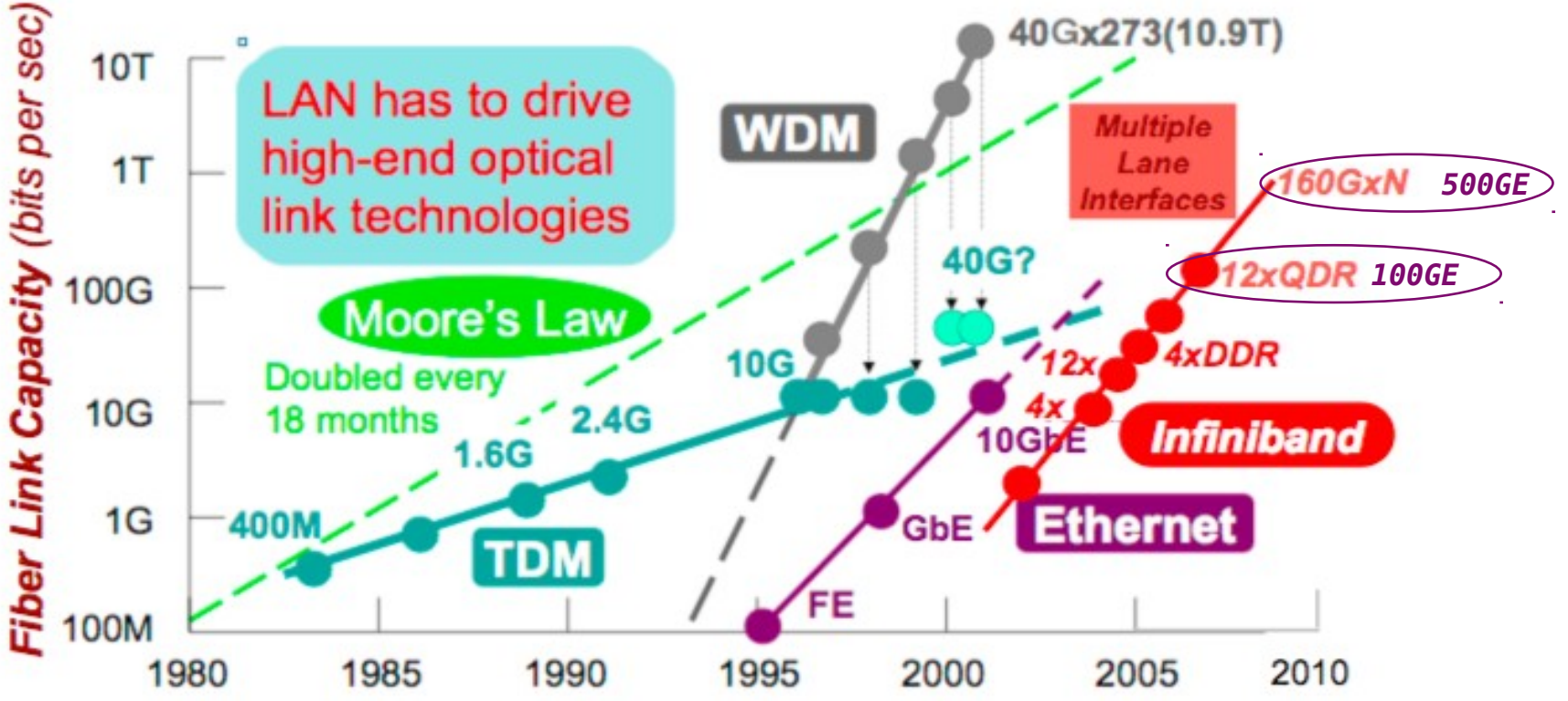
Metcalfe's Law: "The potential value of a network is exponentially proportional to the number of machines connected to it."

Optical Link Performance, per fiber



Ref: O. Ishina, NTT, "Toward Terabit LAN/WAN" Panel, iGrid 2005

Optical Link Performance, per Laser



Ref: O. Ishida, NTT, "Toward Terabit LAN/WAN" Panel, iGrid 2005

Optical Technology Forecast . . .

Yesterday

- *Dispersion Compensation Fiber (DCF) hardwired*
- *No control plan, bandwidth management for lambda services*
- *Static point-to-point optical links, rings and OADMs*
- *Partitioned Access / Metro / LH / ULH optical transport solutions*
- *No layer awareness*

Today

- *Improved economics via large scale photonic integration*
- *Electronic Dispersion Compensation (EDC)*
- *End-to-end GMPLS control plane*
- *Reconfigurable OADMs for wavelength interchange*
- *Dynamic meshed optical nets with flexible Bandwidth Management for wavelength services*
- *Integrated Access/Metro/LH/ULH optical transport solutions*

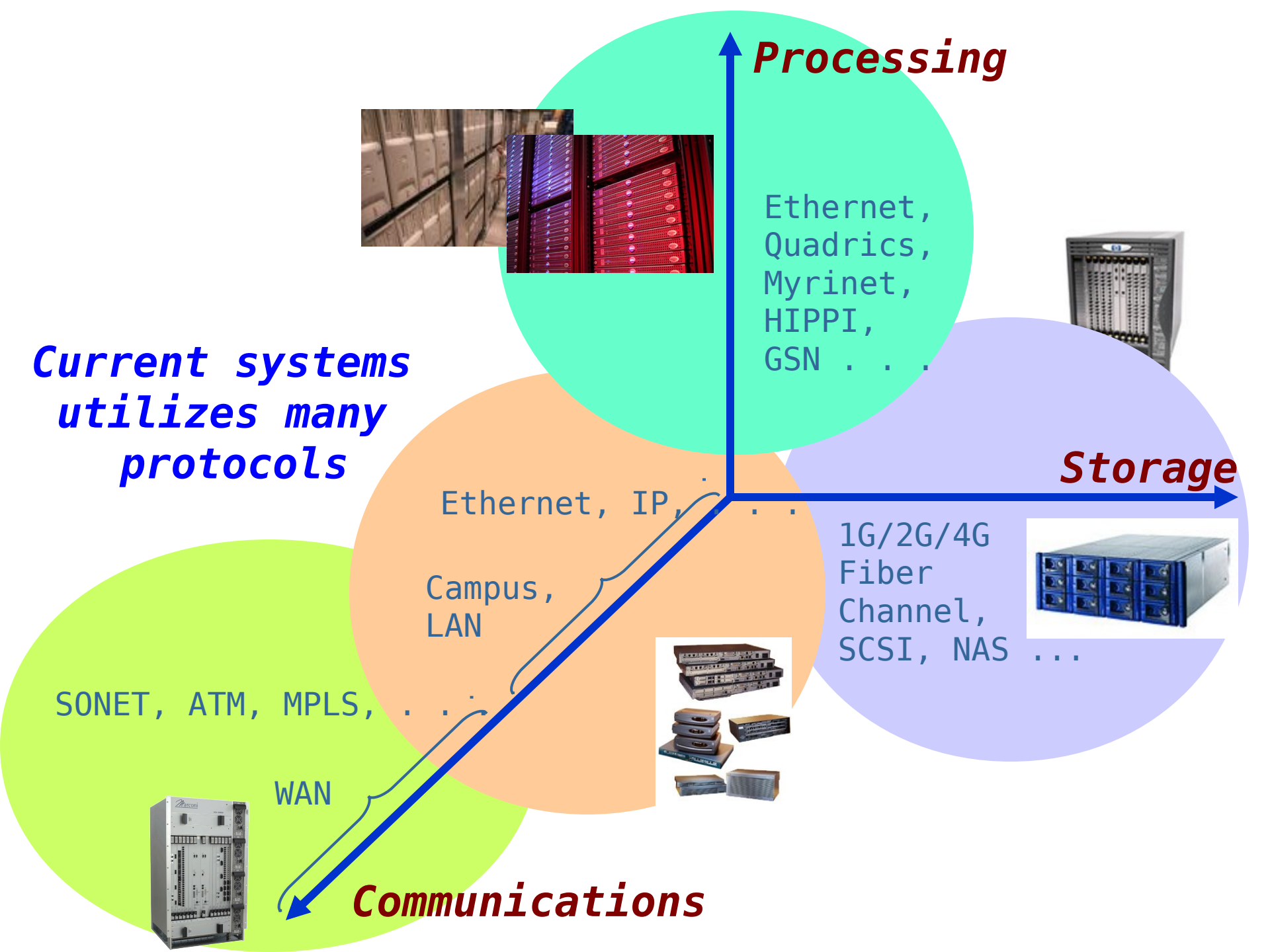
Within 5 Years

- *≥ 1.6 Tbps very large scale photonic integration*
- *≥ 40 Gbps individual channels*
- *≥ 100 Gbps concatenated Channels (Nx100 GbE)*
- *≥ 10 Tbps optical line systems*
- *Multi-domain (L3/L2/L1) IP/DWDM integrated optical netw*

Within 10 Years

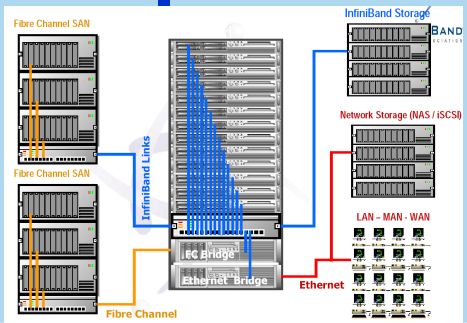
- *≥ 1.0 Tbps single optical flows … switched lambda*

Reference: ONT2 Workshop, NASA Ames, September 2005 … D. Perkins,Infinera & D.


Processing
Ethernet,
Quadrics,
Myrinet,
HIPPI,
GSN . . .
Current systems
utilizes many
protocols
Storage
Ethernet, IP, . . .
1G/2G/4G
Fiber
Channel,
SCSI, NAS ...
Campus,
LAN
SONET, ATM, MPLS, . . .
WAN
Communications

InfiniBand Integrates High Performance Information Systems !

Processing



http://www.infinibandta.org/events/past/it_roadshow/overview.pdf

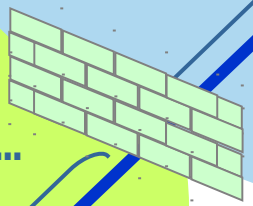
Storage

InfiniBand

- Greater performance,
- Lower latency,
- Easier and faster sharing of data,
- Built in security and
- Quality of Service,
- Improved usability
- Reliability
- Scalability

According to Intel
http://www.intel.com/technology/infiniband/what

Campus



Router Filter Firewalls

tcp

IPV6, ATM, MPLS, …

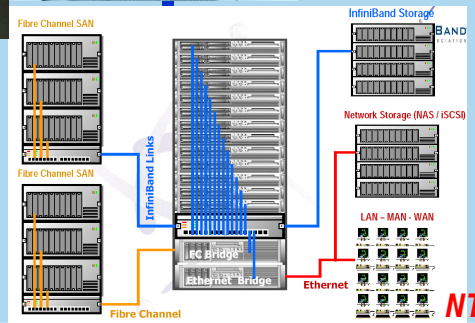
WAN

Communications

Infiniband: A Single Wire Solution

Processing

NTAM



NTAM

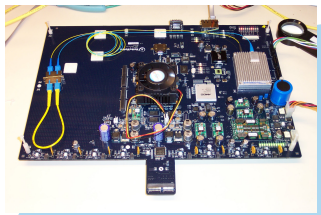
Storage

InfiniBand to WAN Gateway w/ NTAM adds secure WAN to the integrated InfiniBand domain.

InfiniBand

NTAM

http://www.infinibandta.org/events/past/it_roadshow/overview.pdf



Campus

NTAM Firewalls

WAN

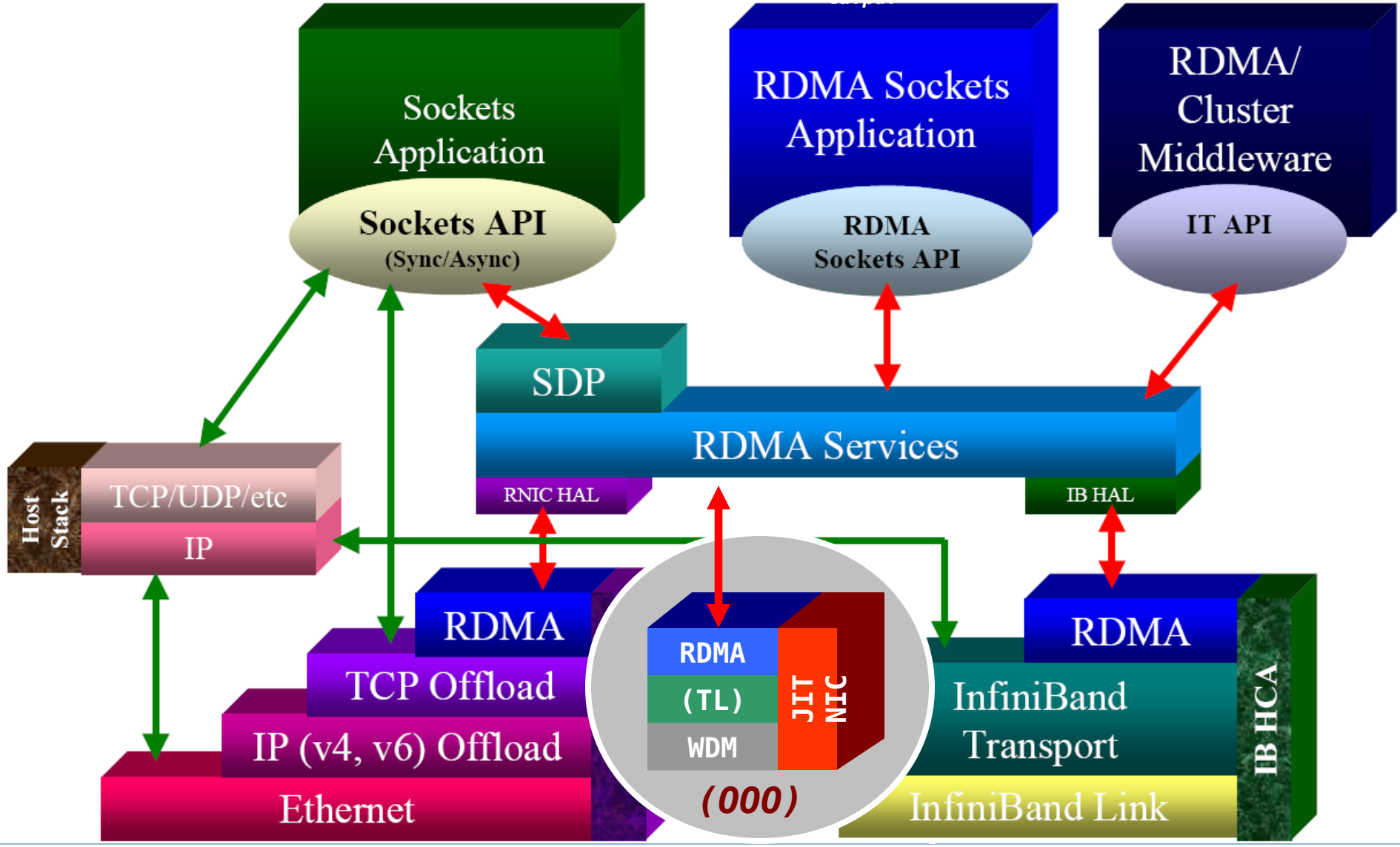
Communications

- Greater performance
- Lower latency
- Easier and faster sharing of data
- Built in security and
- Quality of Service
- Improved usability
- Reliability
- Scalability

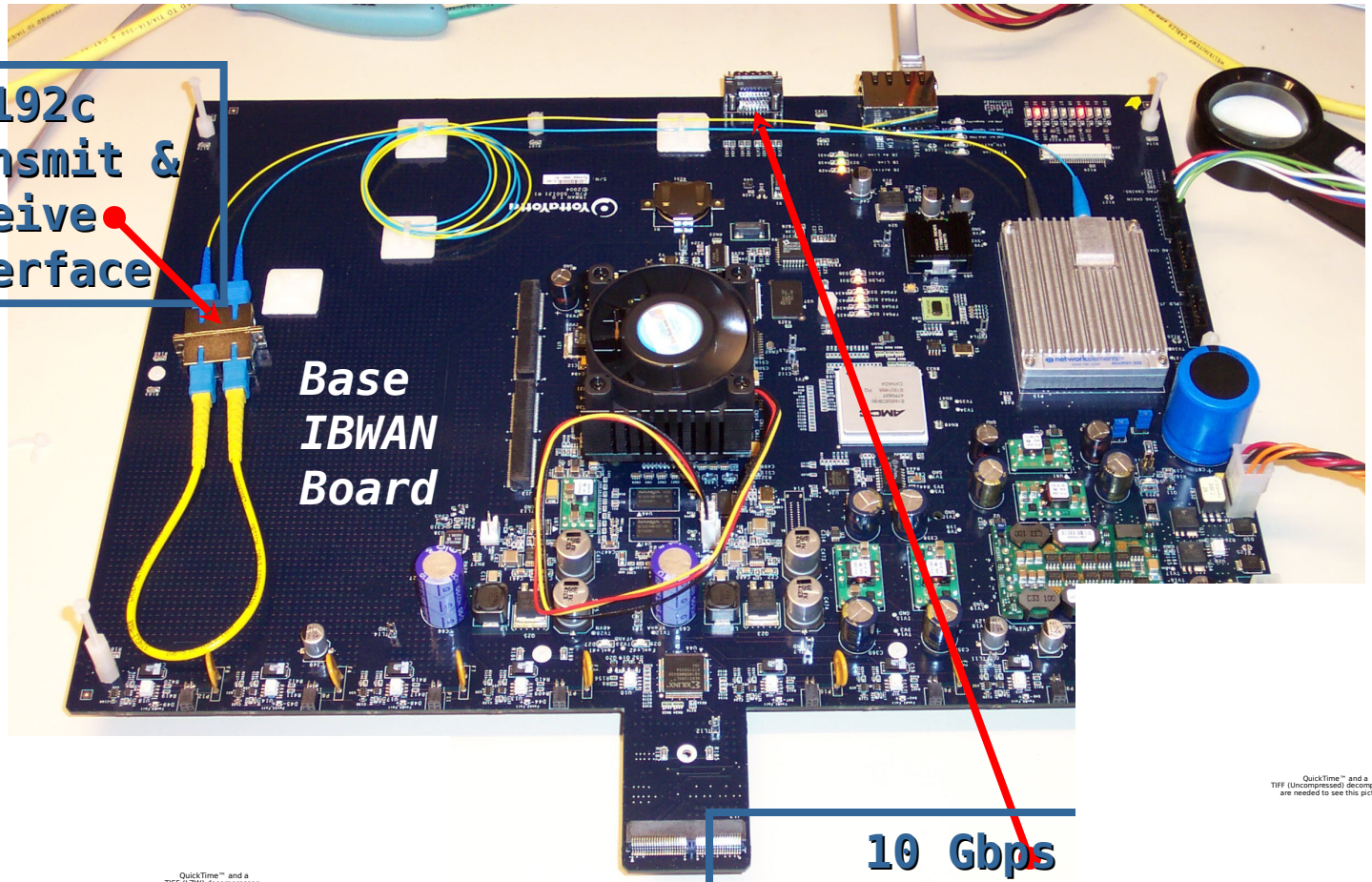
According to Intel
http://www.intel.com/technology/infiniband/whatis.htm

RDMA Infrastructure:
Solution Components

http://www.mellanox.com/shared/hp_ci_oracle_wor



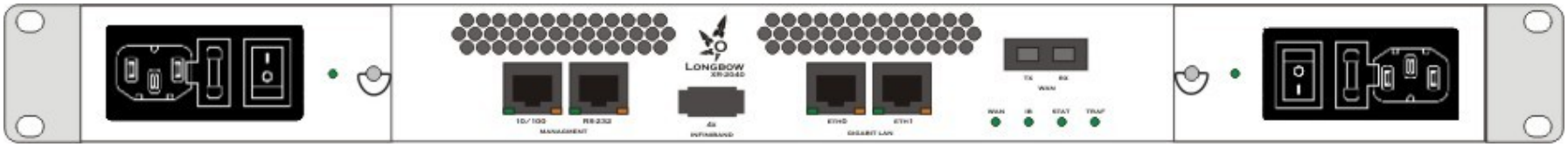
IBWAN: Functional Prototype …



Range Extended InfiniBand . . . Next Step

Performs InfiniBand encapsulation over 10GE, POS and ATM WANs at
4x InfiniBand (10 Gbps, 8b/10b speeds) … *useable w/Type I Encryption*

- *Looks like a 2 port InfiniBand switch or router to the IB fabric*
- *Designed for 100,000 km+ distances for fiber or satcom links*
- *NRL collaborated with Obsidian Research Corp to develop prototypes … flow based, "gargoyle" NTAM sensing, enabling large data streaming is possible in realtime across*
- *Coupled with cache-coherent hardware support from Yotta Yotta*
- *Productized versions of the 10Gbits/s 4xIB prototype*
- *Applications software being developed to facilitate demonstration of wide area switched wavelength IB data streaming*
- *A second source digital hub is available from Bay Micro*



Achieves 950+ MBytes/s sustained performance in a single logical flow ~ 4% CPU load (Opteron 242s using RDMA transport with cache-coherency) … IPv6 Packet Over SONET (for HAIPE when available) & ATM (KG-75a Encryption) modes.

Working toward Terabit Internetworkin

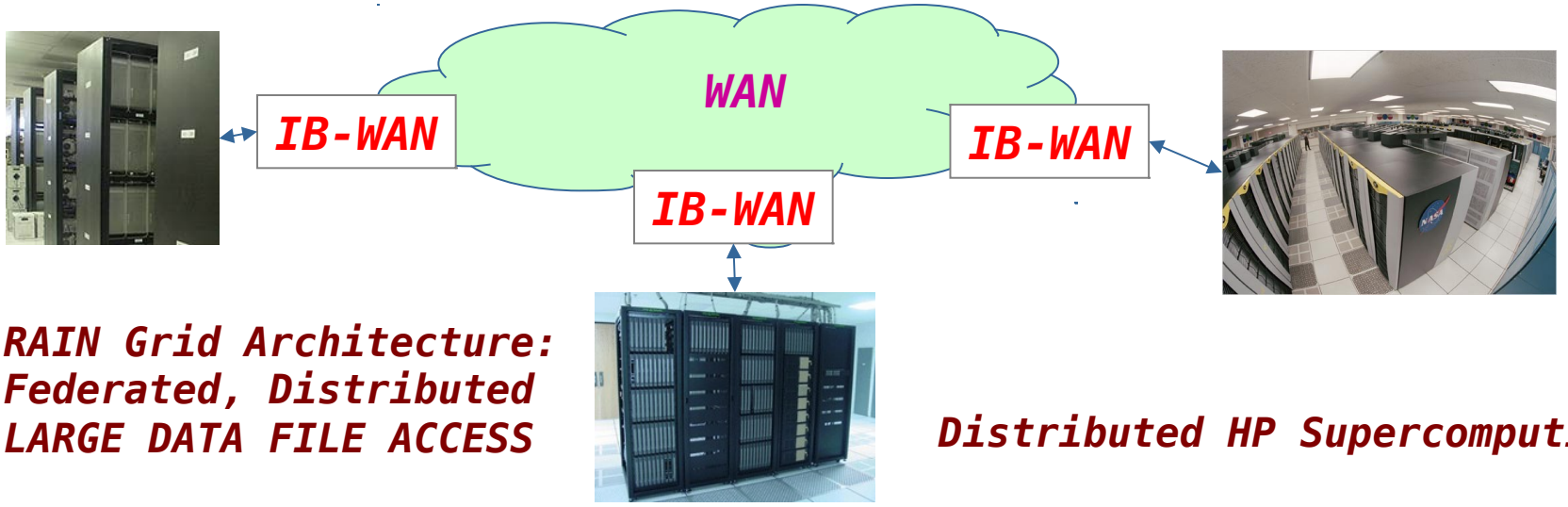
4x IB WAN . . . CY2006

Point-to-point:
- *ATM/SONET (OC-192c)*
- *IPv6 POS (OC-192c)*

Targeted: 3-way multicast
- *ATM with QOS (OC-192c or OC-48c)*
- *IPv6 POS (OC-192c or OC-48c or 10 GigE)*
- *GMPLS (preset)/ JIT (OBS research)*
- *SMPTE 292m (4:2:2 & 4:4:4) 720p/1080p*

12x DDR IB WAN
- *4Q 2006/1Q 2007*
- *GFP*
- *ATM/SONET (OC-768c)*
- *IPv6 POS (OC-768c)*
- *GMPLS (via SIP or UCLP)*
- *JIT (dynamic)*

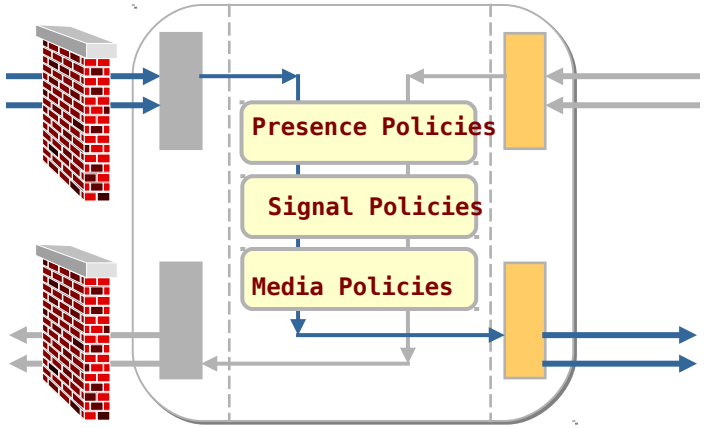
12x QDR
- *~2008 12xQDR=100GE*



RAIN Grid Architecture:
Federated, Distributed
LARGE DATA FILE ACCESS

Distributed HP Supercomputing

SIP: An Application Layer IP Control



- *Realtime Presence Control*
- *Multiprotocol capable*
- *Voice, Video, Data and Imagery flows*
- *Call Signal Control*
- *Media Control*
- *Signaling and Media Encryption*
- *Protocol Validation & Intrusion Protection*
- *Authentication & Authorization*
- *Denial of Service Protection*

Provides a secure fabric for real-time collaboration . . .voice, vide

- *Single view of security & control across enterprise*
- *Enforces corporate, group and user policies: presence, signaling & media*
- *Federation across domains*
- *Hardened appliance*
- *Carrier or Enterprise scalability, availability & security*
- *Utilization of existing infrastructure*
- *Agnostic to transport*

*What is a **GARGOYLE** sensor ?*

Comprehensive Passive Real-Time Flow Monitor

- User Plane and Control Plane Complete Information Assured Transaction Monitoring
- Reporting on System/Network QoS status with every use
  - Capacity, Reachability, Responsiveness, Loss, Jitter
  - ICMP, ECN, Source Quench, DS Byte, TTL

Multiple Flow Strategies

- Layer 2, MPLS, VLAN, IPv4, IPv6, Layer 4 (TCP, IGMP, RTP), 4x/12x IB

Small Footprint

- 200K binary

Performance

- OC-192c, 10GB Ethernet, OC-48c, OC-12c, 100/10 MB Ethernet, SLIP
- *Ongoing research to scaling to OC-768c*
- POS, ATM, Ethernet, FDDI, SLIP, PPP
- > 1.2 Mpkts/sec Dual 2GHz G5 MacOS X.
- > 800Kpkts/sec Dual 2GHz Xeon Linux RH Enterprise

Supporting Multiple OS's

- Linux, Unix, Solaris, IRIX, MacOS X, Windows XP

Comprehensive Data Network Accountability

***NTAM** … Provides an ability to account for all/any network use at a level of abstraction that is useful, all protocols, unencrypted or encrypted, at all layers and for all levels of encapsulation !*

Network Service Functional Assurance

- *Was the network service available?*
- *Was the service request appropriate?*
- *Did the traffic come and go appropriately?*
- *Did it get the treatment it was suppose to receive?*
- *Did the service initiate and terminate in a normal manner?*

Network Control Assurance

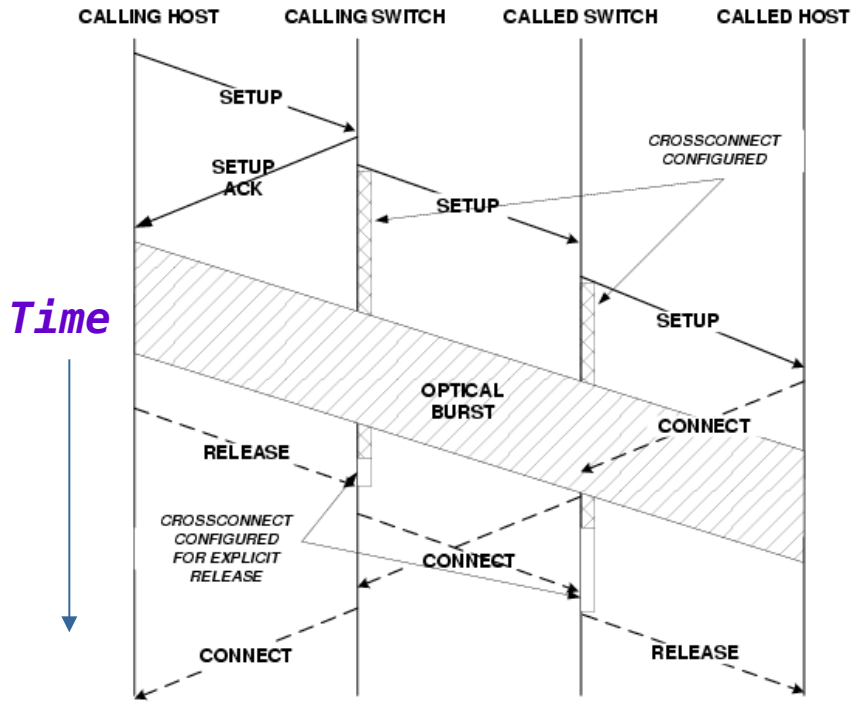
- *Is network control plane operational?*
- *Was the last network shift initiated by the control plane?*
- *Has the routing service converged?*

Information Assurance

- *Converged solution: network, performance, security, billing*

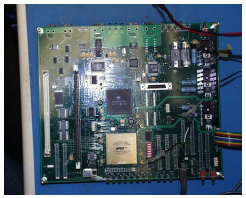
Scalable **Optical Burst Switching … JIT Si**

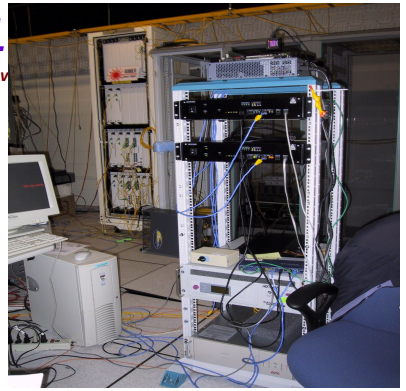
- *No round-trip delay (for 2- or 3-way handshake) required prior to data burst*
- *Out-of-band signaling message precedes data burst*
- *A signaling message's lead time over its data burst shrinks as both propagate through network*
- *Switch resources held only for the duration of burst; no light path required*
- *JIT simplicity -*



Single Burst Exampl

- *Significant improvement in throughput and determinism v TCP/IP/GMPLS*
- *Out-of-band JIT signaling increases communications security & reliability*





Network Scaling Agenda . . .

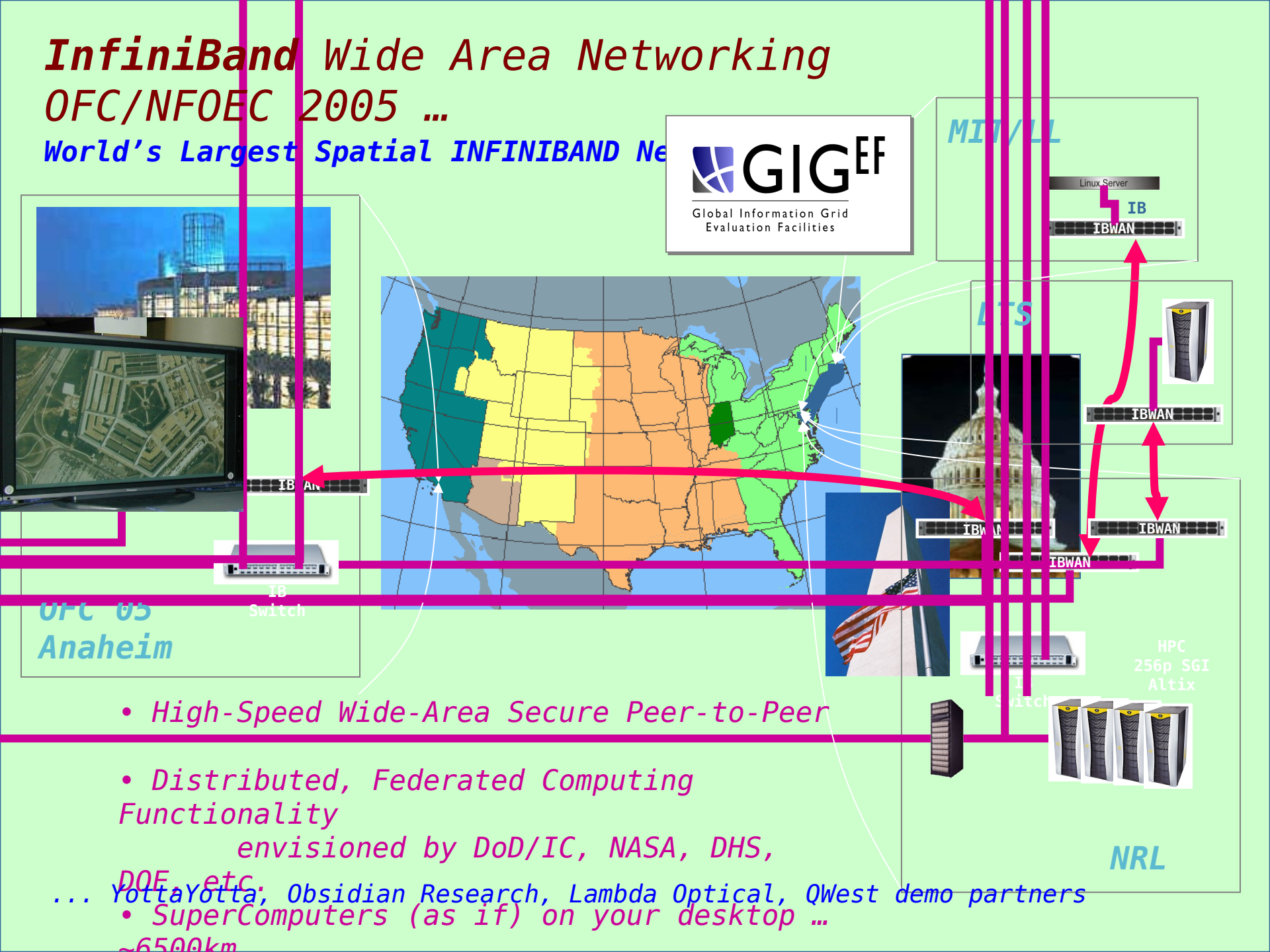
	2005	0-2 YEARS	3-5 YEARS	5-15 YEARS
OPTICAL STREAMS	1-10 Gbps	10-40 Gbps	120-640 Gbps	1-10 Tbps
OPTICAL CNTL Plane	STATIC Provisioned	DYNAMIC (GMPLS)	BURST/JIT Just-in-time	
Control Plane	STATIC Tunnel	DYNAMIC SIP	SIP QoS/QoP	
LAN/WAN Technology	IPV4: 1GE, OC12c, 4xSDR Infiniband	IPV6: 4x/12x SDR/DDR Infbnd(cc), 10GE	IPV6: 12xQDR Infbnd(cc), 100GE, 64-128x IB	All Optical System Interconnect
SECURITY Devices	1.0G IPV4 FW,K5,3DES, CBs, KGs, NTAM	10G KGs, HAIPEs, CAC, FEON, PKI, NTAM	40G HAIPE, Scalable GFP Encrypter	640G HAIPE, GFP Encptr
SPECIAL TOPICS	Quantum Key Distribution (QKD), Dynamic PMD Comp, Peering/Multicast, Parallel Optics, OOO(2R) Optical Regeneration, . . .			

TODAY

InfiniBand Wide Area Networking

OFC/NFOEC 2005 …

World's Largest Spatial INFINIBAND Ne



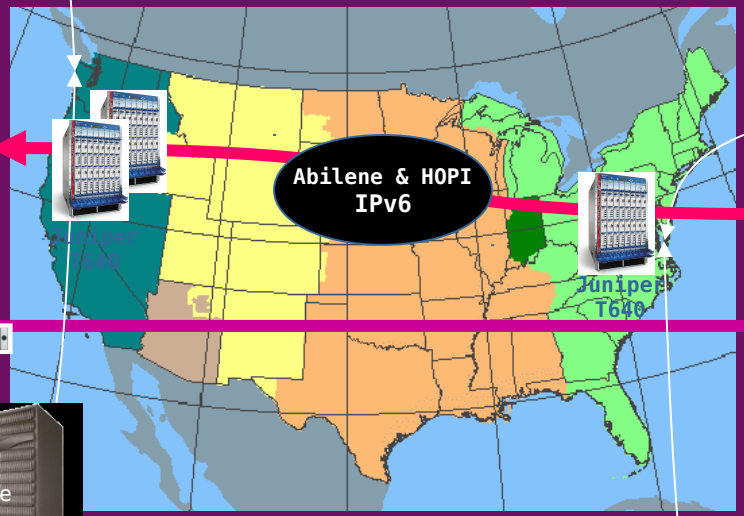
- *High-Speed Wide-Area Secure Peer-to-Peer*
- *Distributed, Federated Computing Functionality envisioned by DoD/IC, NASA, DHS, DOE, etc.*
- *SuperComputers (as if) on your desktop …*

~6500km

… YottaYotta, Obsidian Research, Lambda Optical, QWest demo partners

InfiniBand (IB) Wide Area Networking

SC2005



IBWAN

IB

Voltaire
ISR9024

HD RAVE
Video
retrieval

YottaYotta
GSX3000

SGI
Prism

Voltaire
ISR9024

IB
storage

OPEN IB
ALLIANCE

SC 05
Seattle

IBWAN

Voltaire
ISR9024

YottaYotta
GSX3000

NRL
DC

Voltaire
ISR9096

256p SGI
Altix

SGI
Prism

GIGEF

High-Speed Wide-Area Secure Peer-to-peer
Distributed Computing Functionality needed by
DoD/IC, NASA and DOE
– SuperComputers as if on your desktop

…

SCinet05-InfiniBand 10Gb/s Networ

PEN(B
L I A N C E

Number of Connections

InfiniBand on Fiber
InfiniBand on Copper
InfiniBand on SO
InfiniBand on WI

Lawrence Livermore National Laboratory (CA)

Intel Dupont (WA)

Naval Research Laboratory (VA)

SilverStorm 9120

Voltaire ISR9024

Ciena Corestream

Internet2 Abilene HOPI

Qwest

Cisco ONS15454

Sunnyvale

Ciena Corestream

Pacific NorthWest GigaPOP Westin

Cisco ONS15454

Juniper T640

Cisco 6509

Router to SCinet

GigE

Obsidian Longbow XR

Voltaire ISR9096

SCinet NOC Booth 1316

SilverStorm 9024

SCinet DNOC Booth 6318

SC 05

North Hall

NNSA ASC (LANL, LLNL, SNL) Booth 402

Voltaire ISR9024

SilverStorm 9024

SCinet DNOC Booth 630

6 MPO

StorCloud

SilverStorm 9024

SCinet DNOC Booth 914

2

6

5

SC 05

South Hall

Cisco SFS 7000

SilverStorm 9024

SCinet DNOC Booth 2115

SCinet DNOC Booth 1933

2

Exhibitors

HP – 1608
IBM – 1510
Intel – 1611 (2)

Exhibitor

AMD – 614
SGI – 602
PathScale – 422
PSC – 722
Rackable – 446

Exhibitor

Appro – 703
Cornell – 812
Mellanox – 902
PSC – 722

Exhibitor

Ames Lab – 243
ASUS – 2438
LBNL – 1828
Tyan – 2313

Exhibitor

Cisco – 2218
Dell – 1647
NCSA – 1639

utside

6th

Exhibitors

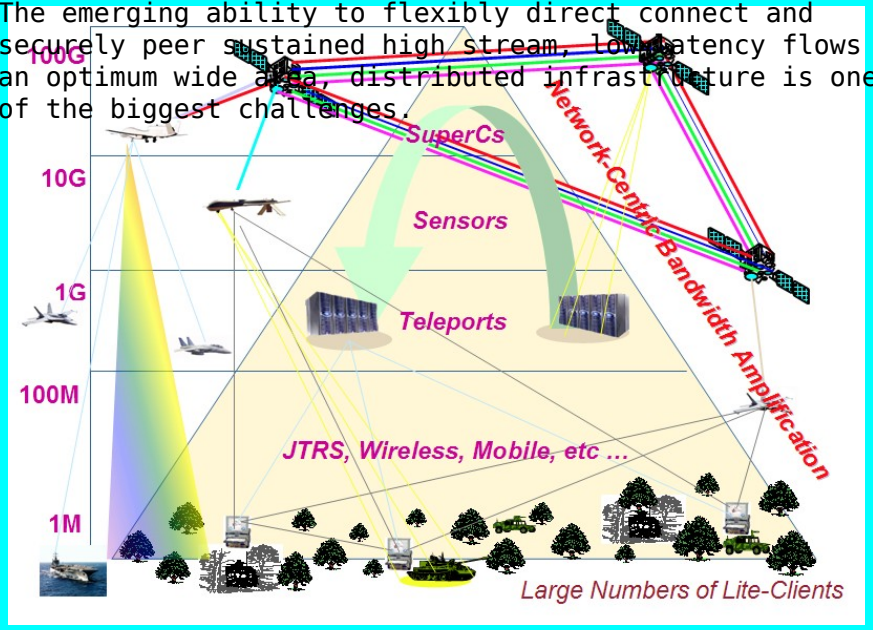
SCALING THE GLOBAL INFORMATION GRID

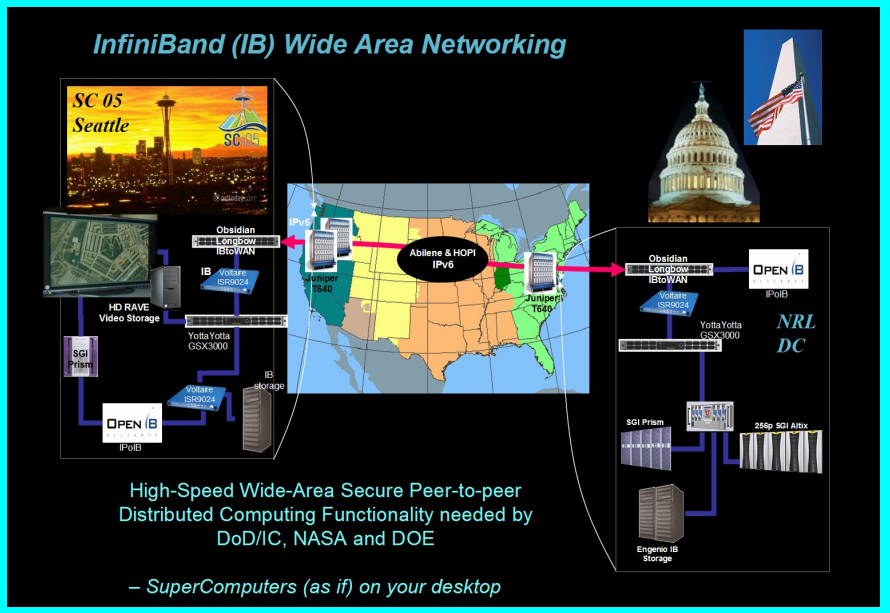
Naval Research Laboratory

Requirements to access and process large amounts of data to exploit information for knowledge have pushed the envelop of conventional architectures. The challenge for High Performance Computing and Communications is to address large data problems in a much more coordinated and rapid manner. This challenge is driven by the exponential growth in data that is driving high-end optical link technology.

Meeting this challenge requires new scalable architectural approaches. Precisely because processing needs to be coupled to distributed, federated global data and the data itself is growing at a rate significantly faster than Moore's Law, a net-centric approach must be employed that meets the conflicting needs of data locality and global consistency. This leads to defining a wholly new edge architecture that can scale to meet the challenges facing the networks in the years ahead.

The emerging ability to flexibly direct connect and securely peer sustained high stream, low latency flows in an optimum wide area, distributed infrastructure is one of the biggest challenges.





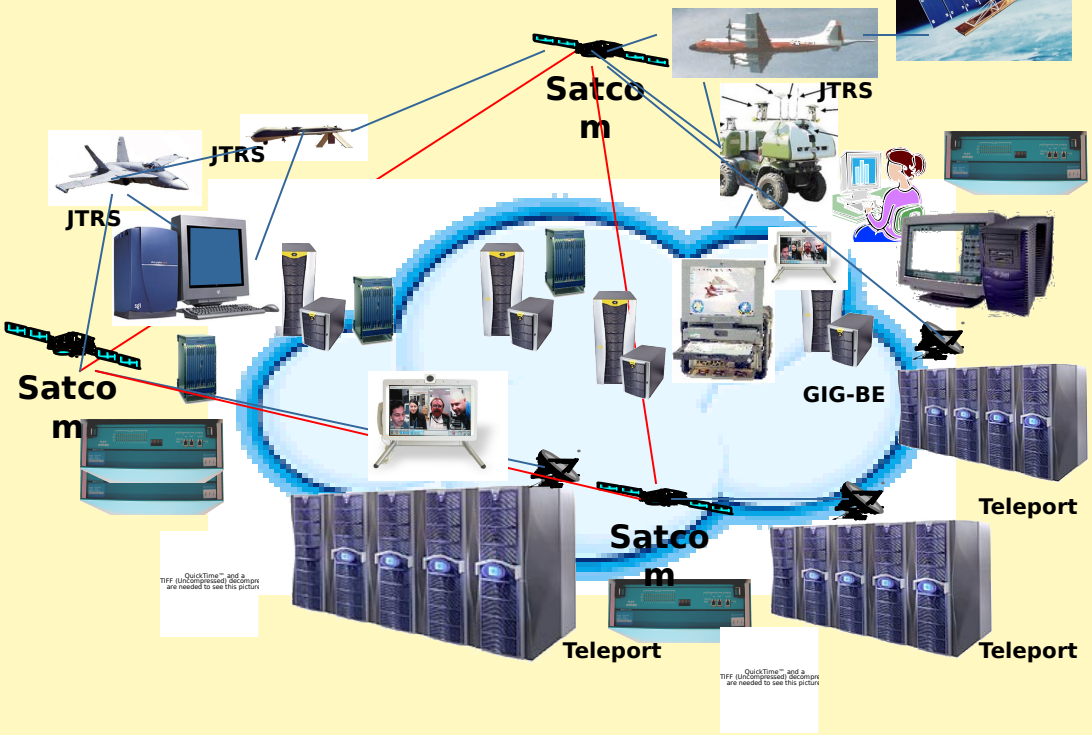
OPERATIONAL REQUIREMENTS

- Global access to the "*right data*" instantly
- Same "right data" everywhere (cache-coherent, synchronized)
- Flexible access for global ***REACHBACK***
- Intuitive access to Large Data Sets (petabytes to exabytes in magnitude)
- Composable remote visualization of large data
- ***TRACEBACK*** for change analysis on an unprecedented scale for signature development, pattern recognition, targeting, forensics, etc.
- ***"Global Information Grid"*** net-centric extension to warfighters deployed or afloat

JCTD: Interactive Distributed Object Library

SOA

❖Virtual network of Active Information Producers & Consumers
… i.e., Grid core w/ P2P edges
❖Vertical fusion - aggregation, delegation
… i.e., level of detail
❖ Horizontal fusion - peer group metadata search & discovery
… e.g., DoD Discovery Metadata Standard
❖Agile data type support for spatiotemporal indexing
❖Pluggable transport architecture including IPV6, native ATM & hardware QoS, DWDM
❖Intelligent caching hierarchy for



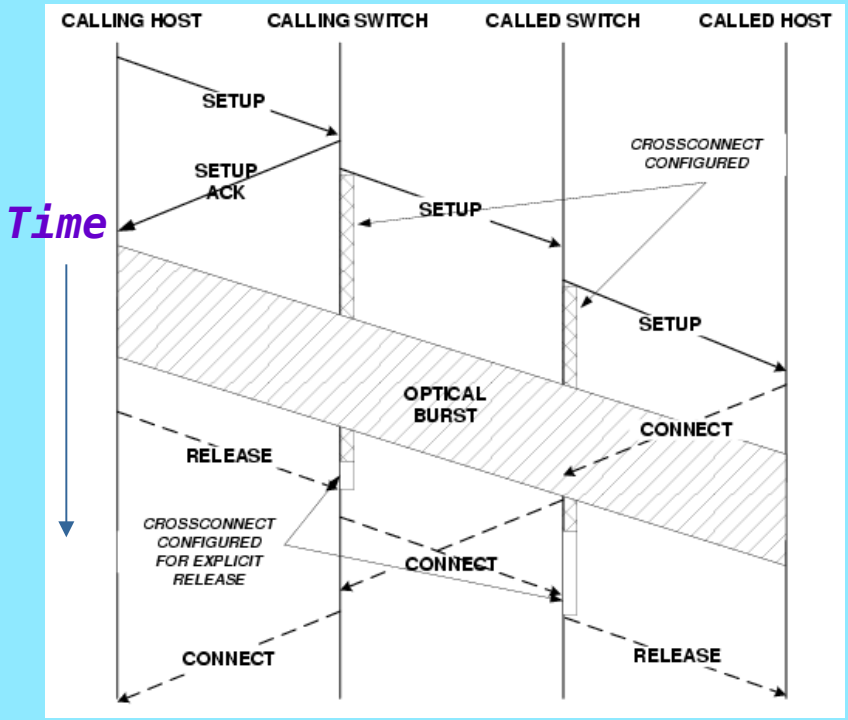
Distributed Database Backend

❖ Immersive Zoomable User Interface (ZUI)
❖ Filter and layer definition, selection, and presentation support
❖ Flexible, intuitive manipulation
❖ Platform support ranging from PDA to workstation to distributed grid to HPCS supercomputer
… High performance: SGI InfiniteReality & UltimateVision systems … well defined API
… Ubiquitous: Desktop PC/Mac/Linux, open sour
… Pervasive: iPAQ handheld

Visualization Front End

Scalable Optical Burst Switching … JIT Si

- No round-trip delay (for 2- or 3-way handshake) required prior to data burst
- Out-of-band signaling message precedes data burst
- A signaling message's lead time over its data burst shrinks as both propagate through network
- Switch resources held only for the duration of burst; no light path required
- JIT simplicity -

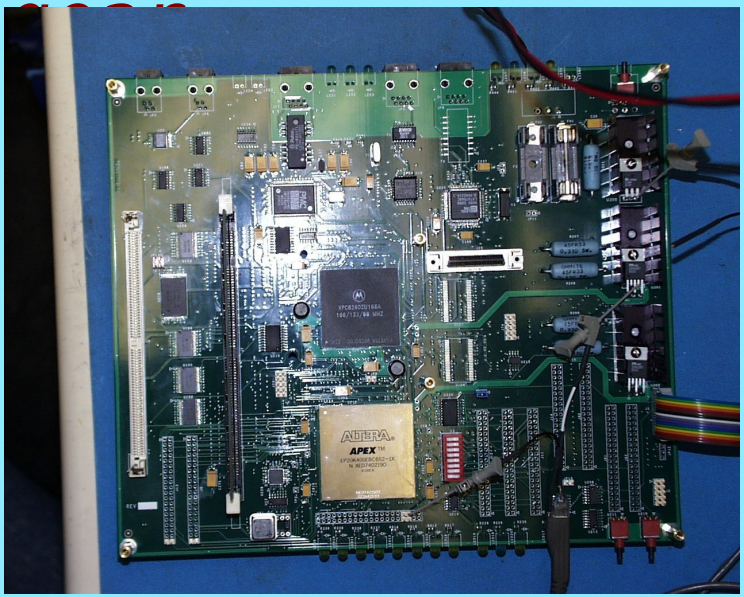


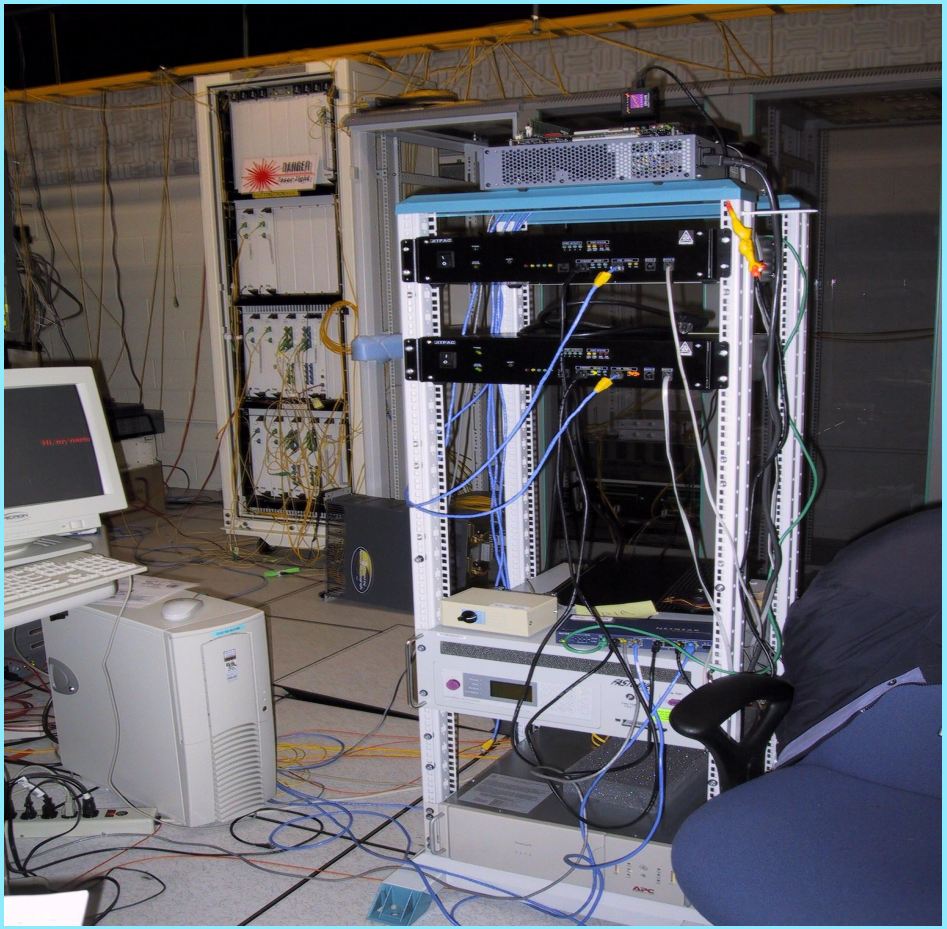
Single Burst Example

- Significant improvement in throughput and determinism vs TCP/IP/GMPLS
- Out-of-band JIT signaling increases communications security & reliability

JITPAC “JIT” NSA Prototype Hardware . . .

NRL Lab with Lambda Optical Systems gear



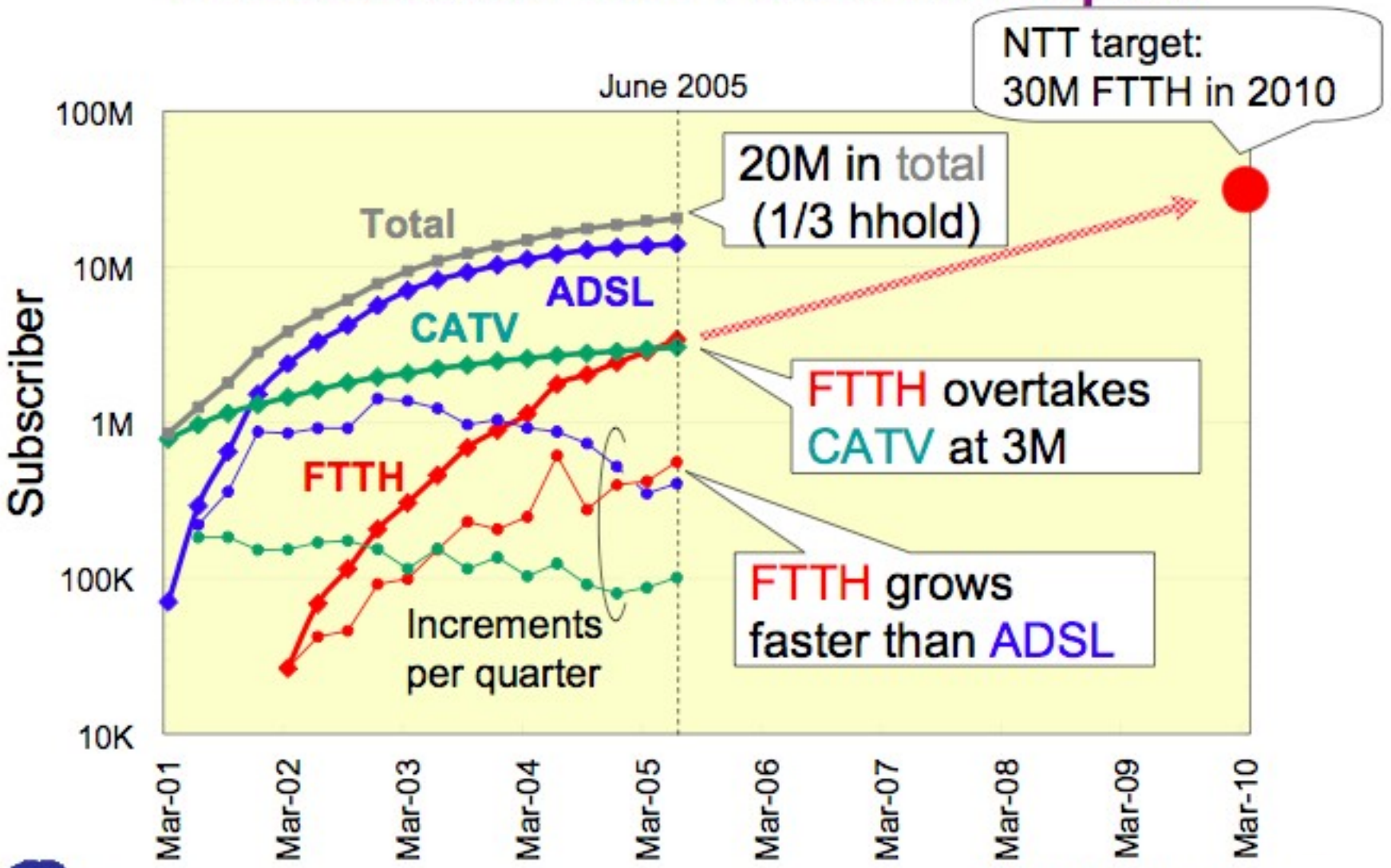


Several KEY Observations . . .

- Large Scale Terabit Optical Testbed(s) are required
  - ***Large data applications will NOT adopt toy infrastructures***
  - ***Simulation is NOT a substitute for the REAL WORLD***
- Infrastructure research is Interdisciplinary

  … ***HPCC, e-Science: Medical, HEP, Visualization, …***
- Infrastructure goals not well understood by Gov't Agencies
- Current Infrastructure Research Funding is Insubstantial
- Japan, Canada, parts Europe, China already

Broadband Services in Japan

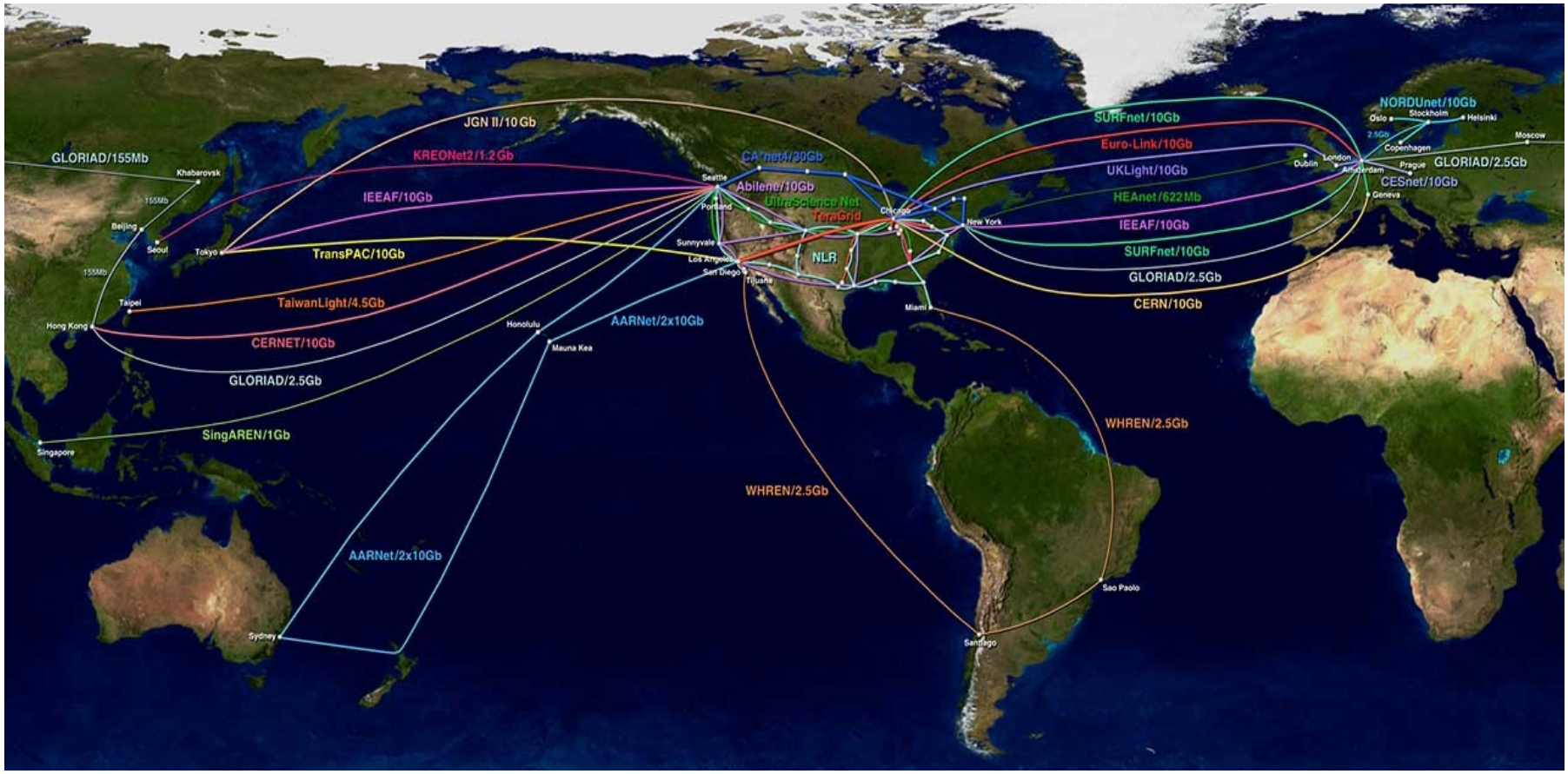


Source: MPT press release

NTT

O. Ishida, "Toward Terabit LAN/WAN", Panel in iGrid 2005

GLIF: Global Lambda Integrated Facil



Visualization courtesy of Bob Patterson, NCSA

SUMMARY . . . Bandwidth with Lowest Latency Most Import

A challenge for net-centric architectures is to provide an *inte*
lowest-latency DISTRIBUTED, FEDERATED INFRASTRUCTURE
that supports moving large data flows globally with QoS/P and Q

Next-gen optical and IP services require a more flexible transp

A powerful set of *new technical capabilities* are under developm
Worldwide GLIF lambda's and Infiniband I/O to meet the challeng
of transporting large data flows with low latency through inter
service grids: RAIN grids for scaled online large data reposito
computational MPI-based grids; visualization grids.

GLIF is a *new process* that has proven effective to instantiate
changes in a short time period and at reasonable cost.

U.S. needs to advance leading edge terabit low-latency flow res
establishing and maintaining a nationwide advanced network infr
to interoperate with GLIF.

"Expose TERABIT interfaces early and ofte

Thank You

nter for Computational Science

of the Naval Research Laboratory